Supplementary Material: A Multimodal Explainable AI Framework for Interpreting Image Classifiers

1 Evaluation of VALE, Multimodal GuISE and S-GuISE XAI approach

Table 1: Paired t-test results for METEOR scores between different XAI methods.

Comparison	***t*-value**	***p*-value**	**Significance**
VALE vs GuISE	-10.064	0.002	Significant ($p < 0.05$)
VALE vs S-GuISE	-3.975	0.028	Significant ($p < 0.05$)
GuISE vs S-GuISE	0.608	0.586	Not Significant

In a paired t-test Table 1, the t-value is a test statistic that quantifies the mean difference between paired measurements compared to the null hypothesis, while the p-value is the probability of observing such a t-value (or more extreme) if the null hypothesis is true. A low p-value (typically ¡ 0.05) indicates statistically significant evidence to reject the null hypothesis, suggesting a real difference between the paired measurements. Paired t-tests were conducted across the same image samples (n=4) to assess statistical significance in METEOR score differences. Paired t-test results indicate that both GuISE (p = 0.002) and S-GuISE (p = 0.028) achieve statistically significant improvements over VALE in METEOR score. The difference between GuISE and S-GuISE is not statistically significant (p = 0.586). These results confirm that our multimodal approach provides quantitatively superior textual explanations compared to VALE.